Does infant-directed speech help phonetic learning? A machine learning investigation

Bogdan Ludusan, Reiko Mazuka, Emmanuel Dupoux

Does infant-directed speech help phonetic learning? A machine learning investigation

Bogdan Ludusan[1,2], Reiko Mazuka[1,3], and Emmanuel Dupoux[4]

[1]Laboratory for Language Development, RIKEN Center for Brain Science
[2]Phonetics Workgroup, Faculty of Lingusitics and Literary Studies, Bielefeld University
[3]Department of Psychology and Neuroscience, Duke University
[4]Laboratoire de Sciences Cognitives et Psycholiguistique, EHESS/ENS/PSL/CNRS/INRIA

A prominent hypothesis holds that by speaking to infants in infant-directed speech (IDS) as opposed to adult-directed speech (ADS), parents help them learn phonetic categories. Specifically, two characteristics of IDS have been claimed to facilitate learning: *hyperarticulation*, which makes the categories more *separable* and *variability*, which makes the generalization more *robust*. Here, we test the separability and robustness of vowel category learning on acoustic representations of speech uttered by Japanese adults in either ADS, IDS (addressed to 18-24 month olds) or read speech (RS). Separability is determined by means of a distance measure computed between the five short vowel categories of Japanese, while robustness is assessed by testing the ability of six different machine learning algorithms trained to classify vowels to generalize on stimuli spoken by a novel speaker in ADS. Using two different speech representations, we find that hyperarticulated speech, in the case of RS, can yield better separability, and that increased between-speaker variability in ADS, can yield, for some algorithms, more robust categories. However, these conclusions do not apply to IDS, which turned out to yield neither more separable nor more robust categories compared to ADS inputs. We discuss the usefulness of machine learning algorithms run on real data to test hypotheses about the functional role of IDS.

Keywords: phonetic learning, speech variability, hyperarticulation, infant-directed speech, adult-directed speech, read speech

Introduction

The way in which infants spontaneously build their phonetic categories from noisy and variable speech input is still a scientific puzzle. A popular, although controversial, hypothesis is that this daunting task is made easier by the fact that parents speak to their children using a special register, called infant-directed speech (IDS). While the characteristics of IDS at the lexical and syntactic levels are, arguably, of a facilitatory nature (*e.g.* Ferguson, 1978), there is still no agreement regarding the helpfulness of IDS for phonetic category learning. One reason for the controversy may be that the phonetic characteristics of IDS are complicated and have been associated to two, somewhat antagonistic, claims.

The first claim is that IDS is a form of *hyperarticulated* speech, whereby the phonetic targets are exaggerated compared to adult-directed speech (although some adult-directed registers also show hyperarticulation characteristics – *e.g.* read speech). For instance, Kuhl et al. (1997) reported increased phonetic distance between the corner vowels (/i/, /a/, /u/) in IDS, in three languages. All other things equal, such expansion of the phonetic space should provide a facilitating effect on learning (see Liu, Kuhl, & Tsao, 2003; Hartman, Ratner, & Newman, 2017; Kalashnikova & Burnham, 2018 for studies showing positive correlations between IDS vowel space measures and language outcome), by mak-

The research reported in this paper was partly funded by JSPS Grant-in-Aid for Scientific Research (16H06319, 20H05617) and MEXT Grant-in-Aid on Innovative Areas #4903 (Co-creative Language Evolution), 17H06382 to R. Mazuka. The work of E. Dupoux in his EHESS role was supported by the European Research Council (ERC-2011-AdG-295810 BOOTPHON) the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute), and CIFAR (Learning in Machines and Brain). Part of the work was conducted while E. Dupoux was a visiting scientist at DeepMind and Facebook. B. Ludusan was also supported by the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement no. 799022.

ing the category means more distant from one another, resulting in better category *separability*. Although the vowel expansion effect has been replicated in other studies (*e.g.* Andruski, Kuhl, & Hayashi, 1999; D. Burnham, Kitamura, & Vollmer-Conna, 2002; Liu et al., 2003, but see *e.g.* Englund & Behne, 2006; Dodane & Al-Tamimi, 2007; Benders, 2013 for a different account), it may not apply to the non-corner vowels (McMurray, Kovack-Lesh, Goodwin, & McEchron, 2013; Cristia & Seidl, 2014), limiting the generality of the putative facilitatory effect.

The second claim is that IDS phonetic categories are more *variable* than adult-directed speech – ADS (*e.g.* Kuhl et al., 1997; de Boer & Kuhl, 2003; McMurray et al., 2013; Miyazawa, Shinya, Martin, Kikuchi, & Mazuka, 2017). This effect is antagonistic to hyperarticulation: While hyperarticulation affects the means of the phonetic categories and makes them more separable, variability affects their standard deviation and makes them more overlapping, hence, less separable.

How do these two effects balance out in practice? One way to test this is to use a measure that combine means and standard deviations (Miyazawa et al., 2017) or measure category discriminability (McMurray et al., 2013; Martin et al., 2015; Guevara-Rukoz et al., 2018). Both types of studies have concluded that the increase in variability for IDS is stronger than the effect of expansion, resulting in a null or slightly negative effect on separability.

However, while recognizing that variability may be detrimental to some aspects of learning (separability), some authors have pointed out that increased phonetic variability could help other learning aspects, such as building more *robust* phonetic categories.

> Mothers addressing infants also increase the variety of exemplars they use, behaving in a way that makes mothers resemble many different talkers, a feature shown to assist category learning in second-language learners. Kuhl (2000), p. 11855.

Considering the counteracting roles of hyperarticulation and variability on phonetic category realization and learning, we aim to investigate here the separability of phonetic categories and the robustness of phonetic category learning, by taking into account the effect of these two phenomena.

Experimental evidence on the impact of variability on robust category learning

Could it be that the detrimental effect of variability for category separability is compensated by increased robustness, once the categories are learned? We review here the adult and infant experimental literature for proof of the impact of variability on robustness in phonetic learning. Solid evidence exists in the learning of non-native phonemic contrasts in *adults* (e.g. /r/-/l/ for Japanese adult learners of English) that phonetic variability during training yields *'robust category formation'* (Lively, Logan, & Pisoni, 1993). This is illustrated by the fact that when trained with multiple speakers, participants can *generalize* the learned contrast to novel words or novel speakers, but not when trained with a single speaker. In this latter case, even though the participants did improve on the training examples, learning failed to generalize to novel speakers. The effectiveness of high variability for phonetic training has been replicated in several studies and is now deployed in practical applications (see a review in Barriuso & Hayes-Harb, 2018). Note, though, that there are at least two differences between these experimental results in adults and the learning situation of infants in their ecological setup.

The first difference is that high-variability studies have focused on between-speaker variability, whereas in the case of IDS, we are dealing with within-speaker variability. These two types of variability could yield different patterns of generalization, although to own knowledge, no adult study has addressed specifically this point.

The second difference is that the aforementioned adult studies trained participants with explicit label categories that were associated to the speech sounds, and they received feedback for their incorrect response, a situation called *supervised learning*. In the case of infants instead, it has been claimed that they learn the categories spontaneously, with weaker or no supervision (unsupervised or self-supervised statistical learning, *e.g.* Kuhl, 2000; Romberg & Saffran, 2010, or using word-level knowledge, *e.g.* Yeung & Werker, 2009; Feldman, Myers, White, Griffiths, & Morgan, 2011). It could be that the effect of variability differs between the former and the latter two types of learning conditions.

Few of the experimental paradigms employed with infants can be compared to those used with adults. Conditioned Head Turning (Kuhl, 1979) trains infants to respond specifically to one class of sounds by turning their head towards it and ignoring the another one. Feedback is provided during the training phase. With this paradigm, Kuhl showed that training a vowel discrimination with stimuli of one speaker can generalize to a different speaker, showing a form of robust response despite low variability input. We are not aware of a study looking at the effect of variability during training with this paradigm. In the Switch paradigm (Werker, Cohen, Lloyd, Casasola, & Stager, 1998), infants are habituated to the pairing between a word and the image of an object, and then tested on their 'surprise' reaction to a mismatch between the word and the image. Assuming the images count as a sort of 'label', this would be similar to supervised training, but without any feedback during training. Rost and McMurray (2009) found that when a single speaker was used during training, infants fail to distinguish between the minimal pairs associated with pictures. Training with a single speaker failed to induce a minimal pair discrimination but training

with multiple speakers succeeded (see also Rost & McMurray, 2010). Other studies showed similar positive effect of variability for the visual referent of words, rather than its phonetic form (*e.g.* Perry, Samuelson, Malloy, & Schiffer, 2010; Gentner & Namy, 1999). In experiments with less supervision, the evidence for a positive effect of phonetic variability is scarcer. For instance, Houston and Jusczyk (2000) used an attention paradigm whereby infants were familiarized with words in isolation and presented passages containing or not these word. They found that 7.5 month-olds would generalize to a novel speaker only if the speaker was of the same gender as the one whose speech was used in the familiarization step, suggesting initial limits to generalization across speakers in early learners. Yet, Houston (2000) found that increasing the variability of speakers during training did facilitate the generalization to novel speakers, consistent with the outcomes of adult high-variability experiments.

To summarize, the experimental evidence regarding the effect of increased variability on phonetic learning is inconclusive, since overall, the strongest evidence of a beneficial effect comes from adult studies with between speaker variability, supervision and feedback during learning. When conditions become closer what infants may experience (within speaker variability, weak or no supervision during learning), the evidence that variability helps becomes scarcer or not available. More generally, while experimental studies in infants and adults are useful in that they point to potential learning effects, the applicability of such effects to real life is limited by the necessarily simplified training regime used during the experiment. Here, we suggest that additional evidence can be obtained through a *computational modelling approach* by asking a slightly different question: does increased variability in IDS help or hinder phonetic category learning *for a particular algorithm*? We intend to address the inconclusiveness of the current state of knowledge by considering both supervised and unsupervised learning models and by separating between within-speaker and between-speaker variability. To the extent that the algorithm is a good model of the infant learner, the results can inform what could happen in infants confronted with similar inputs. Based on the findings of the aforementioned studies we would expect an increased generalizability for supervised models as well as a positive effect of inter-speaker variability on robustness.

Computational studies of the impact of IDS on phonetic learning

Although there is a fairly substantial amount of literature devoted to the computational modelling of phonetic learning (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; Coen, 2006; Vallabha, McClelland, Pons, Werker, & Amano, 2007; Feldman, Griffiths, & Morgan, 2009; Miyazawa, Kikuchi, & Mazuka, 2010; Toscano & McMurray, 2010; Adriaans & Swingley, 2012; Feldman, Griffiths, Goldwater, & Morgan, 2013; Martin, Peperkamp, & Dupoux, 2013; McMurray et al., 2013; Lake, Lee, Glass, & Tenenbaum, 2014; Eaves, Feldman, Griffiths, & Shafto, 2016), only a handful of them have looked at the impact of IDS specifically (de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; Vallabha et al., 2007; Adriaans & Swingley, 2012; McMurray et al., 2013; Eaves et al., 2016).

Kirchhoff and Schimmel (2005) trained an automatic HMM-based word recognition model, using Gaussian mixture components, on IDS and ADS. While no information was given on the age of the infants to which the IDS is addressed, the data contained the same words and was recorded at the same institute as the data used in de Boer & Kuhl, 2003, suggesting they were part of the same dataset (thus, 2-5-month-old infants). They employed Mel Frequency Spectral Coefficients, extracted using a 25 ms window, from several English minimal pair words, and obtained better within-register recognition results for ADS than for IDS, consistent with a negative effect of increased variability in IDS on the separability between categories (see a similar result in McMurray et al., 2013, obtained with a different learning algorithm – logistic regression and different features – the values of the first three formants, on speech addressed to 9-13-month-olds). They also found worse performance when the registers were crossed during training and test. Even though the authors do not present the results in this light, this is actually evidence against the helpfulness of IDS, as the learner is worse off in ADS processing if it was trained in IDS than in ADS. Although both previously mentioned studies compare ADS and IDS learning, they used supervised learning algorithms which, as pointed out above, may not be the best model of the infant learner.

Adriaans and Swingley (2012) employed an unsupervised learning algorithm (based on Expectation Maximization – EM, with a fixed number of Gaussians) on the values of the first two formants of vowels in IDS, and found that the vowels that had acoustic focus (higher pitch, duration or pitch change) yielded better learning than the ones that did not. In this case, data from one mother in a longitudinal study was employed (age of infant between 8 and 14 months). Vallabha et al. (2007) investigated phonetic learning in English and Japanese IDS (addressed to 12-month-old infants), by using two types of algorithms and by taking in input the values of the first two formants and the vowel duration. Both the EM-based learning algorithm, as well as the completely unsupervised one (having no knowledge of the number of phonetic categories it is supposed to learn), similar to Self Organizing Maps, were successful in learning the four categories considered in each of the two investigated languages, although with different performances. While these studies employed models which are more plausible from the point of view of the infant learner than those examined by Kirchhoff and Schimmel (2005) and McMurray et al. (2013), neither of them com-

pared IDS to ADS.

de Boer and Kuhl (2003) used the same unsupervised learning algorithm as Adriaans and Swingley (2012), EM, considering in input the values of the first two formants. The studied IDS data was the one analyzed by Kuhl et al., 1997, containing speech addressed to 2-5-month-old infants. They found that the means of Gaussians trained with IDS speech corresponded more closely to the three corner vowels than Gaussians fitted with ADS. This suggests that IDS provides a better model to learn ADS categories than ADS itself, consistent with the report of exaggerated means in IDS. However, no quantitative analysis of how the IDS-trained Gaussians would actually perform on ADS data was performed.

Eaves et al. (2016) formulated an explicit model of 'teaching', *i.e.* constructing a training sample optimally suited to yield good learning of the ADS categories through an unsupervised learning algorithm (Dirichlet Process Gaussian Mixture Model), based on a parametrization consisting of the values of the first three formants. They noted that the optimal training sample has similar properties to IDS, as reported in the literature (corner vowel hyperarticulation, some non-corner vowel hypoarticulation, increased variability). The study considered both supervised and unsupervised models, tested the generalization to (generated) ADS data, and compared performance obtained on this ADS data with that obtained on their IDS-like distribution. Yet, no quantitative comparisons were made with actual IDS data, the study being conducted with reconstructed, idealized, distributions rather than raw data.

Here again the evidence in favor of the usefulness of IDS for phonetic learning is somewhat mixed, and does not always match the conditions likely to apply to the learning infant. It should also be noted that none of the previous studies have really tested the *robustness* of the phonetic categories, defined as the ability of the system to generalize to a novel, untrained, speaker. We intend to fill this gap, by testing the generalization to novel ADS speakers on actual audio data.

Our approach is presented in Fig. 1. The two axes represent the two main claims regarding the effect of IDS on phonetic learning. The vertical axis represents the claim that hyperarticulated speech yields better separation than standard speech. We test this in Experiment 1 by measuring separability in ADS (which is not hyperarticulated), on the one hand, and IDS and read speech (RS) (which, supposedly, are), on the other. The horizontal axis represents the claim that high variability yields better generalization. We first test this claim for between-speaker variability, in Experiment 2, by manipulating the number of speakers during the learning phase. Then, in Experiment 3, we compare the generalization in ADS and RS (which are considered to have low variability) versus IDS (which has, presumably, high variability). In the following section, we detail and motivate the design choices we made in our computational modelling study.

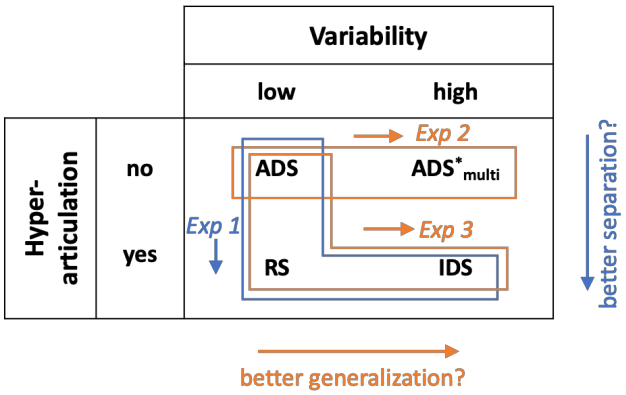


Figure 1. **Hypothesis space tested in this study.** *ADS* and *RS* inputs are assumed to have low acoustic variability, *IDS* and ADS_{multi} (multi-speaker *ADS*) inputs high variability. *RS* and *IDS* are assumed to be hyperarticuled, but not *ADS*. Experiment 1 tests the hypothesis that more hyperarticulation yields better category separation. Experiment 2 and 3 test the hypothesis that more variability yields better generalization. * The variability in ADS_{multi} is between-speaker, in *IDS*, it is within-speaker.

Design choices for the present study

Any computational approach to learning has to specify two key components of the model. The first one is the data used to train the model, and the second one is the learning algorithms used for modelling.

Regarding the data, we use a large and carefully annotated dataset of speech, the RIKEN Mother-Infant Conversation corpus (Mazuka, Igarashi, & Nishikawa, 2006), where the same parents have been recorded in three speech registers: ADS (talking to an experimenter), IDS (playing with or reading a book to their 18-24-month-old toddlers) and read speech (reading a text; RS). The reason we selected this corpus is that it contains high-quality audio recordings of spontaneous infant-directed speech which are entirely manually annotated at the segmental level. In addition, the corpus includes recordings also of ADS and RS from the same mothers, which was crucial for the purpose of the present study – to examine the effect of hyperarticulation and variability on the learnability of phonetic categories. To our knowledge, this is the only dataset that allows a direct comparison among IDS, ADS, and RS by the same speakers. A series of studies have already established that Japanese IDS presents the main characteristics of IDS documented in various languages (Ferguson, 1964): shorter sentences, repeated words, and exaggerated intonation (*e.g.* Fernald et al., 1989; Andruski et al., 1999; Amano, Nakatani, & Kondo, 2006), while exhibiting also language-specific properties, such as different vocabulary structure (Fernald & Morikawa, 1993). The IDS characteristics shared with other languages

were found also for the age range (18-24 months) present in the RIKEN corpus, in particular that IDS has higher pitch (Igarashi, Nishikawa, Tanaka, & Mazuka, 2013), shorter utterances (Martin, Igarashi, Jincho, & Mazuka, 2016), and an expanded vowel space (Miyazawa et al., 2017) compared to ADS. We focus here on the five short vowels of Japanese, which enables comparison with other computational modelling work on phonetic learning (*e.g.* Vallabha et al., 2007; McMurray, Aslin, & Toscano, 2009, among others). Moreover, as evidence from other languages shows no change in the size of the mother's vowel space with the age of the addressee (and, this, for a larger age range, overlapping the one present in the RIKEN corpus – Liu, Tsao, & Kuhl, 2009; E. Burnham et al., 2015), we would expect similar findings also for different age ranges.

Regarding the algorithm, there is a wide variety of views regarding how infants achieve phonetic learning, and each of these views can be implemented in a variety of ways, yielding different algorithms. Computational studies have typically used a narrow range of such algorithms, making it difficult to know whether the results are general or specific to the chosen algorithms. Here, our strategy was to (1) cover some of the most popular algorithms used in previous studies (for comparability), and (2), organize them systematically in terms of their basic assumptions (for interpretability). We sorted the algorithms according to two dimensions.

The first one regards the amount of innate constraints or inductive biases that the learning model brings to the task. Many studies have used parametric algorithms, which assume that the underlying phonetic categories have particular shapes, typically, Gaussian distributions over the input dimensions (see de Boer & Kuhl, 2003; Vallabha et al., 2007; McMurray et al., 2009). Other studies have used non-parametric algorithms, which make no such assumptions and can accommodate categories of different shapes (such as the self organizing maps of Kohonen, 1988, as in Gauthier, Shi, & Xu, 2007 or in Vallabha et al., 2007).

The second dimension relates to the amount of top-down information available to infants. At one extreme, supervised models assume that the learner is presented for each speech instance with a category label. This is the case for many second language learning paradigms where adults are taught to label or discriminate non-native speech sounds (Lively et al., 1993). We consider this a control condition, as it is highly implausible that infants have access to such systematic information. At the other extreme, unsupervised algorithms assume that the learner has no top-down information at all: just the speech input. This hypothesis has been proposed under the name of "distributional learning", and has been tested both in infants with artificial categories (Maye, Werker, & Gerken, 2002) and in models with more or less natural speech inputs (Vallabha et al., 2007; McMurray et al., 2009). In between these two cases, there is a continuum of algorithms that use varying amount of top-down information in the shape of already learned words (Jansen & Niyogi, 2007; Thiolliere, Dunbar, Synnaeve, Versteegh, & Dupoux, 2015) or algorithms performing joint word and category learning (Feldman et al., 2009). Here, we chose a class of algorithms which is basically unsupervised, but uses top-down knowledge to inform the number of speech categories to be found (see Fourtassi, Schatz, Varadarajan, & Dupoux, 2014 for a mechanism for finding such numbers without assuming perfect word segmentation nor perfect word categorization). Such class of algorithm were used in previous computational studies (de Boer & Kuhl, 2003; Adriaans & Swingley, 2012).

These two dimensions are crossed in a factorial design, resulting in six different machine learning algorithms (three parametric: Naive Bayes, Expectation Maximization, Dirichlet Process Gaussian Mixture Model, and three non-parametric: Nearest Neighbour, Hierarchical Clustering, Self Organizing Maps; two supervised: Naive Bayes, Nearest Neighbour, two partially unsupervised: Expectation Maximization, Hierarchical Clustering, and two unsupervised: Dirichlet Process Gaussian Mixture Model, Self Organizing Maps), as displayed in Table 1. Since Expectation Maximization and Dirichlet Process Gaussian Mixture Model have been previously employed for phonetic learning modelling, we made use of them in this study. Moreover, we chose the supervised algorithm of the same class (the underlying probabilistic model being as in the other two, a mixture of Gaussian) most similar to them, Naive Bayes. Similarly, Self Organizing Maps have been employed in modelling studies and we chose for the other two non-parametric models the simplest algorithms that could represent this particular crossing of factors. Thus, Nearest Neighbour and Hierarchical Clustering, two algorithms based on the distance between points, with no underlying assumption about the shape of the categories, were chosen.

Finally, for computational modelling, it is important to note that the input (the type of information taken to represent each input token) to the algorithm may matter almost as much as the algorithm itself. We run our six algorithms on two commonly used input representations: high-level language-specific parameters (the values of the first two formants, as in Coen, 2006; Vallabha et al., 2007; McMurray et al., 2009; Adriaans & Swingley, 2012) and low-level acoustic features derived from spectrograms and used in speech recognition (Mel Filter Cepstrum Coeficients – MFCC, as in de Boer & Kuhl, 2003; Kirchhoff & Schimmel, 2005; Miyazawa et al., 2010; Martin et al., 2016; Guevara-Rukoz et al., 2018).

To sum up, we ran six learning algorithm (two levels of inductive biases, three levels of supervision) crossed by two input representations (formants and MFCCs), on the short Japanese vowels spoken in three registers (ADS, IDS, and RS). In Experiment 1, we verify that our stimuli have similar

Table 1
Summary properties of the 6 algorithms in this study. [1] *de Boer and Kuhl (2003); Adriaans and Swingley (2012); Eaves et al. (2016).* [2] *Feldman et al. (2009); Eaves et al. (2016); close variants were used in Vallabha et al. (2007); Toscano and McMurray (2010); Lake et al. (2014)* [3] *Coen (2006); Vallabha et al. (2007); Miyazawa et al. (2010)*

	Known	Known no. of	Type of Model	
Type of Supervision	labels?	categories?	Gaussian	Non-Gaussian
Supervised	yes	yes	Naive Bayes (NB)	Nearest Neighbour (NN)
Partially Unsupervised	no	yes	Expectation Maximization (EM)[1]	Hierarchical Clustering (HC)
Unsupervised	no	no	Dirichlet Process Gaussian Mixture Model (DPGMM)[2]	Self Organizing Maps (SOM)[3]

characteristics regarding hyperarticulation and variability to those described previously for IDS, ADS and RS, as well as the effect of these two phenomena on the phonetic category separability. That is, we predict that IDS should both be more hyperarticulated and more variable than ADS, whereas RS should be more hyperarticulated, but probably less variable than ADS. With respect to separability, taking into account previous work, we expect a small negative or no effect on IDS, compared to ADS. In Experiment 2 we investigate the claim that variability can help generalization by manipulating the number of speakers present in the training data. We train our six algorithms on two speech representations, using ADS data, and we test the generalizing to novel ADS speakers. Based on the findings of previous studies, one would expect that training on data from multiple speakers would give a better generalization than training on a single speaker only. In Experiment 3, we test how each speech register, at training time, helps generalizing to novel ADS speakers, at test time. If IDS variability somehow mimics speaker variability and if this helps generalization, we expect IDS training to yield better performance than ADS training. The predictions regarding RS are less straightforward. On the one hand, the higher degree of hyperarticulation would suggest that RS training will yield good category learning. On the other hand, if variability helps robustness, we predict less robust categories after RS training, to the extent that RS is indeed less variable than ADS.

Experiment 1

In this experiment, we conduct two sets of analyses. The first set investigates the two properties attributed to IDS categories, namely hyperarticulation and variability. Based on past work (Miyazawa et al., 2017), we expect IDS to be both hyperarticulated and more variable than ADS. RS, in contrast, should be hyperarticulated, but less variable than ADS. Compared to the analysis conducted in Miyazawa et al. (2017), we employ two new metrics, relying on F1-F2 measures and low-level spectral features, respectively. We operationalize hyperarticulation as the average distance between category centers, and variability as the average distance within category. Even though the speech corpus is the same as in the Miyazawa et al. (2017) study, it is important to check that our particular selection of stimuli shows the expected characteristics.

The second set of analyses tests the effect of register on the separability of vowel categories. Separability, or its converse – category overlap, is a function of both hyperarticulation (helpful) and variability (detrimental). Miyazawa et al. (2017) used an inter-class distance, which assumes a parametric shape to the categories, to measure separability, while Martin et al. (2015) used the machine ABX discrimination score, which is non-parametric. Both found that IDS was less separable than ADS. Here, we adopt the method employed by Miyazawa et al. (2017), computing the same distance between vowel classes. Based on the previous studies, we expect that IDS vowels to be harder to separate than ADS vowels, with RS vowels being the most separable.

Methods

Dataset. The data used in this study belongs to the RIKEN Mother-Infant Conversation corpus (Mazuka et al., 2006). The corpus consists of speech uttered by 22 Japanese mothers to their 18-24 month-old toddlers, while interacting with them either through the use of toys or by reading a book. The same mothers have been recorded also talking to an adult experimenter, about topics related to child rearing. The resulting datasets contain over 11 hours of infant directed speech and around 3 hours of adult-directed speech. Besides the IDS and ADS recordings, we also considered a third dataset, comprising read speech. Twenty out of the total of 22 mothers in the RIKEN corpus were recorded reading a set of sentences having the same phoneme distribution as Japanese ADS. The recordings contained in this corpus can be seen as a more formal and carefully pronounced speech register, further called RS (read speech) in this paper. All three datasets have been fully transcribed and annotated at the segmental level.

We computed the number of occurrences of each of the five Japanese short vowels (/a/, /e/, /i/, /o/, /ɯ/), for each speaker in our three datasets. Then, we considered only the speakers which had, for each of the five vowel categories and across the three registers, at least 100 vowel instances,

resulting in 15 speakers. For these speakers, we randomly selected, from each register and vowel category, a number of examples equal to the minimum number of examples in any vowel class (107). Thus, our final dataset had, for each register and speaker, 5×107 vowel instances, totaling 24,075 vowel tokens.

As high-level representations, we used the first two formants (F1 and F2) values of each vowel, obtained using Praat (Boersma, 2002), a software for phonetic analyses. The first five formants were extracted, considering as maximum value for the formant search range 5500 Hz and applying pre-emphasis to frequencies above 50 Hz. The values of the first two formants extracted from the center of each 25 ms analysis frame (with a 10 ms frame shift) were employed. Thus, for each frame we had a feature vector composed of two values.

Additionally, low-level audio representations (Mel Frequency Cepstral Coefficients – MFCCs) were extracted from each vowel. They were computed as follows. Fist, a short term power spectrum was computed every 10 ms over a window of 25 ms (modelling the frequency decomposition in the cochlea). The different frequencies were then averaged over a mel scale (corresponding to the auditory critical bands), and a logarithmic compression was applied (reducing the dynamic range). The resulting log spectrum was converted back into the time domain via a discrete cosine transform, and only the first 12 coefficients (plus the signal energy) were retained. Up to and including the log compression, these steps are similar to those used in models of auditory processing. The discrete cosine transform is a technique to make the different components statistically independent, and is similar to running a principal component analysis over the log spectrum. We employed the Python package spectral[1] for the extraction of the MFCC features.

Analysis. In order to determine the hyperarticulation and variability measures, we represent each vowel by the feature values (two formants or thirteen MFCCs) extracted from the central frame of the vowel. For calculating the hyperarticulation, we define the category center as being its centroid (the vowel closest, on average, to all the other vowels of that category) and the average Euclidean distance between category centers is reported. Variability is computed as the average Euclidean distance between all vowel pairs from the same category.

To compute separability, we calculated the normalized Euclidean distance between each vowel class pair, for each register and speaker separately, employing the same measure as in Miyazawa et al. (2017). The distance between two vowel classes i and j is defined in Equation 1, where K represents the size of the feature vector, μ_{ik} the mean and σ_{ik} the standard deviation of the kth element of the feature vector for the class i. It represents the distance between the means of the classes, normalized by their standard deviation. Thus, for equal mean values, a lower standard deviation would return a higher distance. Similarly to the hyperarticulation and variability measures, the distance used for determining separability was computed on the features extracted from the central frame belonging to the selected vowels.

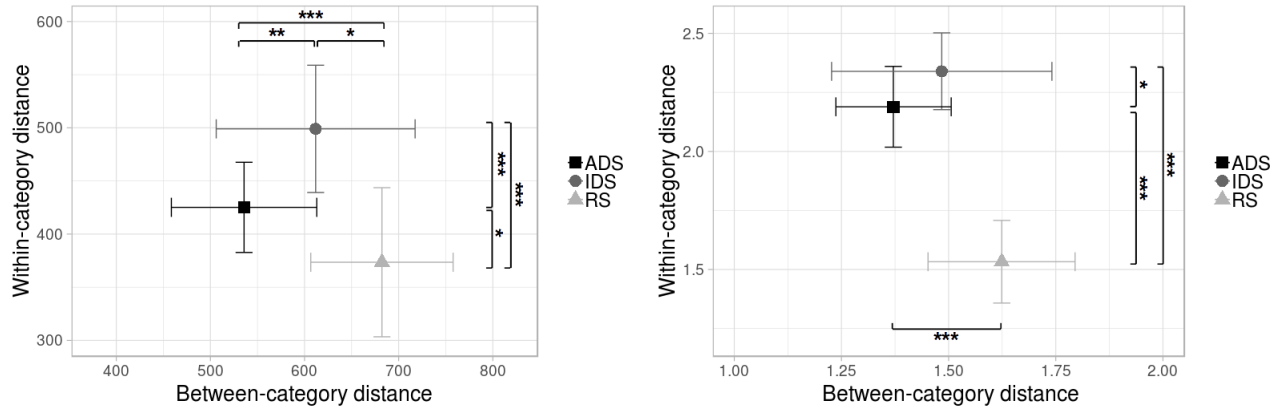
$$D_{ij} = \sqrt{\frac{K \sum_{k=1}^{K} (\mu_{ik} - \mu_{jk})^2}{\sum_{k=1}^{K} \sigma_{ik}^2 + \sum_{k=1}^{K} \sigma_{jk}^2}} \quad (1)$$

For each of these measures, ANOVAs and paired two-tailed t-tests were applied to the per-speaker results in order to check the statistical significance of the differences between registers.

Results and discussion

We illustrate in Fig. 2 the results for hyperarticulation and variability (see Fig. I of the Supplementary Materials for a more detailed illustration, showing individual speaker values). A two-way ANOVA, with distance as dependent variable, register as independent variable, and speaker as random variable was run, separately, for each type of distance (between-category for hyperarticulation and within-category for variability). For the formant features, it showed significant register effects for both variability [$F(2, 42) = 17.37, p = 3.2e^{-6}, \eta^2 = .453$] and hyperarticulation [$F(2, 42) = 10.57, p = 1.9e^{-4}, \eta^2 = .335$]. Analysing hyperarticulation with t-tests revealed a similar pattern to the one reported by Miyazawa et al. (2017): more hyperarticulation for RS than for ADS ($t = -6.01, df = 14, p = 3.2e^{-5}$), for IDS than for ADS ($t = -3.40, df = 14, p = .004$), as well as for RS than for IDS ($t = -2.51, df = 14, p = .025$). We then investigated the hypearticulation of the three corner vowels (defined as the average of the /a/-/i/, /a/-/ɯ/ and /i/-/ɯ/ distances) observing, also in this case, a larger distance between these categories in IDS than in ADS ($t = -4.14, df = 14, p = .001$). However the hyperarticulation effect observed for the corner vowel categories was not more enhanced, compared to the overall hyperarticulation ($t = 1.99, df = 14, p = .066$). In terms of variability, our results further replicate the findings of Miyazawa et al. (2017): a high variability for IDS, followed by ADS (ADS-IDS: $t = -4.22, df = 14, p = 8.6e^{-4}$) and the lowest variability for RS (ADS-RS: $t = 2.37, df = 14, p = .033$, IDS-RS: $t = 7.33, df = 14, p = 3.7e^{-6}$). We then examined whether the age of the infant has an effect on the hyperarticulation or variability present in IDS. For this, we fitted two linear regression models, with the infant's age (in days) as continuous independent variable. None of the ANOVA analyses performed on the models showed a significant effect of age on our variables of interest ([$F(1, 13) = 1.14, p = .305, \eta^2 = .081$] for hyperarticulation and [$F(1, 13) = 2.37, p = .148, \eta^2 = .154$] for variability).

[1] https://github.com/mwv/spectral



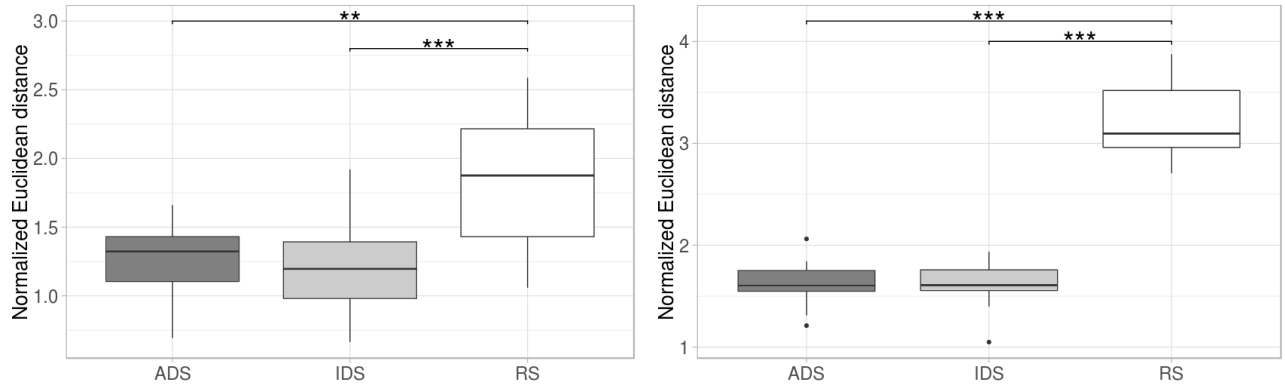
(a) Formants (b) MFCCs

Figure 2. Between-category distance (Hyperarticulation) versus Within-category distance (Variability), averaged across the five Japanese vowels and across speakers, for formant features (a) and MFCCs (b). Displayed are the p-values of uncorrected paired *t*-tests (*: $p < .05$, **: $p < .01$, ***: $p < .001$).

For MFCCs, the ANOVA revealed significant register effects for both variability [$F(2, 42) = 96.16, p < 2e^{-16}, \eta^2 = .821$] and hyperarticulation [$F(2, 42) = 6.34, p = .004, \eta^2 = .232$]. Post-hoc *t*-tests revealed a similar pattern to the one obtained with formant features: more hyperarticulation for RS than for ADS ($t = -5.89, df = 14, p = 3.9e^{-5}$) and for IDS than for ADS, but no difference between IDS and RS ($t = -1.76, df = 14, p = .1$). Although the ADS-IDS contrast was found to be only marginally significant ($t = -1.95, df = 14, p = .071$), no previous study that found hyperarticulation in IDS looked at the entire speech spectrum. It might be that the phonetic enhancement is limited to or more pronounced in the lower part of the spectrum, where the first two formants are found. When looking at the average distance between the three corner vowel, we see no difference in hyperarticulation between ADS and IDS ($t = -1.44, df = 14, p = .172$) and no difference between the hypearticulation of all the vowels versus that of the corner vowels ($t = -0.36, df = 14, p = .724$). Also for variability, the results are similar to the ones attained with formants: the highest variability for IDS, followed by ADS and then RS ($t = -2.8, df = 14, p = .014$ for ADS-IDS, $t = -9.82, df = 14, p = 1.2e^{-7}$ for ADS-RS and $t = 14.43, df = 14, p = 8.5e^{-10}$ for IDS-RS). Repeating the same analyses as for formant features, revealed no significant effect of the infant's age on either hyperarticulation ([$F(1, 13) = 0.74, p = .406, \eta^2 = .054$]) or variability ([$F(1, 13) = 0.47, p = .504, \eta^2 = .035$]).

Turning now to the results obtained for separability, a two-way ANOVA analysis, with the normalized Euclidean distance as dependent variable and register as independent variable, showed a significant effect of register [$F(2, 42) = 13.55, p = 2.9e^{-5}, \eta^2 = .392$], for formant features. The best separability was obtained with the RS dataset, followed by ADS and IDS. Post-hoc *t*-tests showed that the difference in separability between ADS and IDS was not significant ($t = -0.38, df = 14, p = .71$), whereas the difference between RS and ADS was significant ($t = -3.75, df = 14, p = .002$) (the IDS-RS difference was significant: $t = -5.63, df = 14, p = 6.2e^{-5}$). An identical ANOVA analysis revealed a significant effect of register [$F(2, 42) = 168.7, p < 2e^{-16}, \eta^2 = .889$] also when spectral representations were employed. Similarly to the formant feature set, the best separability was reached with the RS dataset, followed by ADS and IDS (ADS-IDS: $t = 0.07, df = 14, p = .94$, ADS-RS: $t = -16.32, df = 14, p = 1.7e^{-10}$, IDS-RS: $t = -13.18, df = 14, p = 2.8e^{-9}$). No effect of infant's age on separability was observed for either formants ([$F(1, 13) = 0.47, p = .147, \eta^2 = .155$]) or MFCCs ([$F(1, 13) = 0.47, p = .368, \eta^2 = .063$]). Fig. 3 shows the results obtained for separability, employing the two feature sets.

These findings are in line with those one would expect based on the results for within- and between-category distance: RS, being both hyperarticulated and not very variable, yields the best separability results. IDS tends to be hyperarticulated, but also more variable, with no overall positive effect on separability (an ANOVA restricted to ADS and IDS revealed no effect of register: [$F(1, 28) = 0.13, p = 0.72, \eta^2 = .005$] for formants and [$F(1, 28) = 0.006, p = 0.94, \eta^2 = .0002$] for MFCCs), results which are congruent with those of Miyazawa et al. (2017). The counteracting effects of hyperarticulation and variability for IDS compared to ADS and the net lack of positive effect on separability is also similar to Guevara-Rukoz et al. (2018), despite a very different analysis method. Overall, RS comes across as an unequivocal case of clear speech, with both hyperarticulation and reduced variability, giving a much better separability than the other two registers. IDS, on the other hand, does not



(a) Formants (b) MFCCs

Figure 3. Normalized Euclidean distance (within-speaker separability) for the five Japanese short vowels, assessed in three registers (ADS, IDS, RS), for formant features (a) and MFCCs (b). The computed distance is averaged across all vowel pairs of a speaker. Displayed are the p-values of uncorrected paired *t*-tests (*: $p < .05$, **: $p < .01$, ***: $p < .001$).

qualify as clear speech, at least as far as learning of phonetic categories is concerned.

Experiment 2

In this experiment, we test the role of between-speaker variability on the learning of phonetic categories and its generalization to a new speaker in ADS. For this, we train each of the algorithms to categorize the five Japanese vowels on a subset of the tokens (the training set) and compute the classification error rate on a different subset (the held out test set). We consider two conditions: in the mono-speaker condition the 'infant' is trained with a single 'parent', while in the multi-speaker case, the training is done with a large 'family' of 14 speakers. In both case, the amount of exposure (number of vowel tokens) is kept the same. If robustness is helped by variability, one should obtain better generalization in a large, rather than in a small 'family setting'. In both conditions, the testing is always done on a novel speaker.

Methods

Learning algorithms. Six different learning algorithms were employed in this study: Naive Bayes (NB), Nearest Neighbour (NN), Expectation Maximization clustering (EM), Hierarchical Clustering (HC), Dirichlet Process Gaussian Mixture Model (DPGMM), and Self Organizing Maps (SOM). The first two are supervised (assuming that the category labels are available during training), the last two are unsupervised (no information besides the speech representation) and the middle two we call partially unsupervised, as the only supervision comes from knowing the number of phonetic categories. Half of these algorithms (NB, EM, DPGMM) assume that categories are Gaussian, half do not. For the first five algorithms we used the implementation offered by the scikit-learn machine learning library (Pedregosa et al., 2011), while the last algorithm was part of the SOMbrero package (Villa-Vialaneix et al., 2018).

The NB algorithm is a probabilistic parametric supervised classifier which assumes that each input feature is independent from one another and follows a different Gaussian distribution given a class value. In other words, the categories are assumed to be Gaussians with a diagonal covariance matrix, whose optimal parameters are estimated by the classifier during training. At test time, the posterior probability of each class is computed by decomposing it using Bayes' formula and predicting the class label having the highest probability.

The NN classifier is an instance-based non-parametric supervised learning method. It does not assume that the categories have any particular shape. Instead of deriving statistics from the training example, it stores each training example (with their class label) and uses them directly at prediction time. At test time, the algorithm computes the Euclidean distance between the given instance and the instances stored during training, and assigns to the new instance the same class as its closest training instance.

The EM algorithm employs an unsupervised parametric learning paradigm. It makes the same assumptions about the shape of the categories as NB, but does not use any class label at training time. It tries to fit n Gaussian distributions to the training data, by means of the Expectation Maximization algorithm, where n is the expected number of categories. At test time, the algorithm will return the probability of each instance of belonging to each of the clusters. The system was given the number of vowel classes, five, and it was run for a maximum of 100 iterations, with a convergence threshold of $1E-3$ and using full covariance matrices.

HC is an unsupervised non-parametric method which builds a hierarchy of clusters. We employ here the 'bottom-

up' approach, also called agglomerative clustering, in which each observation starts in its own cluster. Then, moving up in the hierarchy, new clusters are created by merging existing ones, such that the sum of squared differences within the clusters is minimized. Since this method creates a tree-structure based on the training data, in order to be able to predict cluster labels for unseen data, we use the predicted labels on the training set to train a nearest neighbour classifier. Thus, for each test instance, the classifier will return the cluster label of the closest (in terms of Euclidean distance) training observation. The number of clusters, five, was given as a parameter to the model.

DPGMM is an unsupervised Bayesian learning method. It represents an extension of Gaussian Mixture Model used in the EM algorithm, as it allows an infinite number of components, while being able to automatically determine the number of clusters from the data. Here, we limit the number of components to 37, with each component using a full covariance matrix. The number 37 was chosen as it represents the maximum number of vowels (excluding long and nasal vowels) that can be produced by the vocal apparatus, as illustrated by the IPA vowel chart (*IPA Chart*, 2015) and including the diacritic vowels. The model was run for a maximum of 2000 iterations, with a convergence threshold of $1E-3$. The hyperparameters were set to their default values, since it has been previously shown that they have a reduced effect on phoneme class learning (Chen, Leung, Xie, Ma, & Li, 2015).

The SOM algorithm is an unsupervised non-parametric method based on artificial neural networks, that uses competitive learning to map the input space into a lower dimensional representation. This representation has the property that more similar observations are mapped closer together than less similar observations. We used a 6×6 grid (giving 36 nodes, similar to the number of components used for DPGMM, 37) with a square topology. The algorithm was run for a maximum of 500 iterations, employing a Gaussian neighbourhood with a Euclidean distance and a hard affectation.

Analysis. To compute generalization, all six learning algorithms were run using the same experimental setting: The sampled vowel instances, for each register and speaker, were randomly split into a train and a test set, respectively, with the train set containing 87 instances of each vowel and the test set the remaining 20 instances. The models were trained and tested separately for each of the 3 registers and 15 speakers. The training and test sets were used in a mismatched condition (*e.g.* testing on one speaker while having trained on another speaker). In the mono-speaker case, 210 tests (14 train speakers x 15 test speakers) were run using ADS data, averaged within speaker and then the average across the speakers reported. For the multi-speaker case, a train set was created for each speaker, containing randomly sampled vowels from the mono-speaker condition train sets of all the other 14 speaker, except the one on which we tested. The distribution of speakers was uniform, while keeping the amount of training instances constant (87 instances x 5 vowels). Thus, we tested on one ADS speaker, while having trained on a set containing vowels from the remaining 14 speakers and we computed the average across the 15 speakers.

The same feature vectors were used to represent each speech frame as in the previous experiment (F1/F2 values or 13 MFCCs). Differently from the analysis in Experiment 1, which considered only the central frame of each vowel to compute hyperarticulation, variability and separability, we use here all the frames of a vowel. For example, if a vowel has a length of 15 frames, the machine learning algorithms will classify each of the 15 frames individually. As the classifiers take a frame-based decision, returning class-probabilities (or binary 0/1 decision values, in the case of NN, HC and SOM) for each frame, and wanting to perform per-phoneme evaluations, we summed the class probabilities across all frames belonging to a vowel instance and the class having the highest sum was considered to be the predicted one. For the evaluation of the unsupervised algorithms (EM, HC, DPGMM and SOM), the obtained clusters were first mapped to the five phoneme classes, by minimizing the classification error on the training set, and then the same evaluation as for supervised methods was applied. The results were evaluated using the F-score, a standard evaluation measure for classification tasks. It represents the harmonic mean of precision (the proportion of correctly classified instances out of the total number of instances classified as belonging to that class) and recall (the proportion of correctly classified instances out of the total number of instances belonging to that class). It takes values between zero and one, the latter value representing a perfect classification. Because the unsupervised algorithms (DPGMM and SOM) may classify vowels into a sixth class (containing all instances not being assigned to one of the five gold classes), we use the micro-averaged F-score, which computes the true positives, false positives and false negatives over the entire five classes. Each speaker contributed one data point to the statistical analyses, the classification F-score obtained for the test set corresponding to that speaker.

Results and discussion

The results obtained are illustrated in Fig. 4.[2] We investigated the role of number of speakers in the train set condition, the type of feature used, the supervision type and the inductive bias of the models, by fitting a linear model with these factors as independent variables and the classification F-score as the dependent variable. A subsequent ANOVA analysis revealed significant main effects of the following predictors: supervision type [$F(2, 336) = 288.6, p < 2.2e^{-16}, \eta^2 = .432$], inductive bias [$F(1, 336) = 65.4, p = 1.1e^{-14}, \eta^2 = .049$] and feature type [$F(1, 336) = 146.7, p < 2.2e^{-16}, \eta^2 = .110$], as well as significant interactions between supervision type and feature type [$F(2, 336) = 75.3, p < 2.2e^{-16}, \eta^2 = .113$] and between supervision type, inductive bias and feature type [$F(2, 336) = 23.1, p = 4.1e^{-10}, \eta^2 = .034$]. The two-way interaction between number of speakers and supervision type [$F(2, 336) = 2.3, p = .099, \eta^2 = .003$], as well as between number of speakers and inductive bias [$F(2, 336) = 3.4, p = .067, \eta^2 = .003$], were found marginally significant.

The analysis shows that increasing the number of speakers present in the training set has little or no effect on the overall generalizability. For formant features, the effect is not uniform across classes of algorithms, partially unsupervised algorithms being helped by an increased number of speakers in the train set (a post-hoc *t*-test showed a significant effect for both EM: $t = -3.59, df = 14, p = .003$, and HC: $t = -2.44, df = 14, p = .029$), and little change for the other classes of algorithms. For MFCCs, the direction of the effect depends on the class of algorithms, a positive one for supervised algorithms (only NB reached significance, $t = -9.72, df = 14, p = 1.3e^{-7}$), and a negative one for unsupervised algorithms (only SOM significant, $t = 2.43, df = 14, p = .029$).

This result is interesting. It shows that the idea that high variability is beneficial to induce robust learning is not logically warranted. Only when the number of phonetic categories is known (supervised and partially unsupervised algorithms) does a higher number of speakers bring a small significant improvement in the learning performance. In contrast, when the learning algorithm is completely unsupervised, high variability has either no effect or a detrimental one. The net result of speaker variability is therefore dependant on the learning strategy employed by infants. A pure bottom-up infant would be hurt by high variability, but an infant relying on some sort of lexical feedback might gain some benefit from it (to the extent that the high variability does not, itself, impede lexical learning).

Experiment 3

In this experiment, we test whether exposure to the phonetic variability of IDS could help build more robust categories, that generalize to a new speaker in ADS.[3] We compare this to a putative infant who would be trained on ADS or RS. If robustness is helped by IDS variability, one should obtain better generalization when training on this register, than with ADS or RS data. The tests are conducted on each speaker separately, resulting in a generalization score for each speaker.

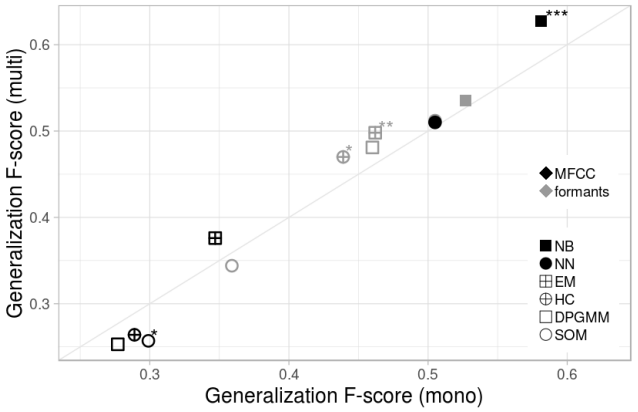


Figure 4. Within-register generalization to novel ADS speakers for the classification of the five Japanese short vowels by six learning algorithms, trained with ADS data from one speaker (mono) or 14 speakers (multi), on formant features and MFCCs. The scores represent average F-scores cross-validated on a held out test set of one novel speaker. Displayed next to each point are the p-values of uncorrected paired *t*-tests (*: $p < .05$, **: $p < .01$, ***: $p < .001$). The grey line represents the equal-performance line. Note that the points corresponding to NN, for the two feature types, are overlapping.

Methods

The same format of train and test sets were used as in the mono-speaker condition of Experiment 2. Separate train sets were created for each register and speaker, while the test sets were identical as in the previous experiment (ADS data). Then, for each register, 210 tests were run, within-speaker average classification performance computed and the average across the 15 speakers presented. The same evaluation

[2] We report here the results obtained when running each algorithm once, with a more detailed analysis included in the Supplementary Materials, Section S2. While three of the employed models (EM, DPGMM, SOM) are stochastic, the deviation across 100 runs was low (in the majority of cases, it was lower than 0.005, with a maximum obtained deviation of 0.016 for the DPGMM MFCC multi case). An analysis of the results taking into account all 100 runs is presented in the same section of the Supplementary Materials.

[3] For an experiment testing the generalizability to a novel IDS speaker, the reader is invited to see Section S4 of the Supplementary Materials.

was performed as previously.

Results and discussion

Fig. 5 illustrates the detailed findings, with the ADS-IDS comparison in the left panel and the ADS-RS comparison in the right panel.[4] The interaction between register, the characteristics of the learning algorithms (type of supervision and inductive bias) and the feature type were analyzed by fitting a linear model with F-score as the response variable and the previously mentioned variables as predictors. An ANOVA analysis of the model fitted with the ADS-IDS data revealed significant main effects of all the predictors: register [$F(1,336) = 17.2, p = 4.2e^{-5}, \eta^2 = .011$], supervision type [$F(2,336) = 376.5, p < 2.2e^{-16}, \eta^2 = .467$], inductive bias [$F(1,336) = 81.5, p < 2.2e^{-16}, \eta^2 = .051$] and feature type [$F(1,336) = 192.4, p < 2.2e^{-16}, \eta^2 = .119$], as well as significant interactions between register and supervision type [$F(2,336) = 4.6, p = .011, \eta^2 = .006$], between supervision type and feature type [$F(2,336) = 90.8, p < 2.2e^{-16}, \eta^2 = .113$] and between supervision type, inductive bias and feature type [$F(2,336) = 11.9, p = 9.7e^{-6}, \eta^2 = .015$]. The four-way interaction of the predictors was found marginally significant [$F(2,336) = 2.9, p = .054, \eta^2 = .004$].

Running the same analysis on the ADS-RS data, showed similar results to the ADS-IDS comparison: a significant main effects of all the predictors, register [$F(1,336) = 14.2, p = 2.0e^{-4}, \eta^2 = .012$], supervision type [$F(2,336) = 277.8, p < 2.2e^{-16}, \eta^2 = .458$], inductive bias [$F(1,336) = 35.3, p = 7.1e^{-9}, \eta^2 = .029$] and feature type [$F(1,336) = 113.0, p < 2.2e^{-16}, \eta^2 = .093$], as well as significant interactions between register and supervision type [$F(2,336) = 5.0, p = .007, \eta^2 = .008$], between supervision type and feature type [$F(2,336) = 50.4, p < 2.2e^{-16}, \eta^2 = .083$], between inductive bias and feature type [$F(1,336) = 5.7, p = .018, \eta^2 = .005$]. and between supervision type, inductive bias and feature type [$F(2,336) = 15.4, p = 3.9e^{-7}, \eta^2 = .025$].

The previous analyses show that the higher variability present in IDS does not make it good for generalization. We can actually see that in a majority of cases, IDS trained models are worse than ADS trained models. In fact, RS, which is less variable than ADS, manages to yield better generalization than ADS itself, both in the case of formant features (for two of the six algorithms) as well as in the case of spectral representation (for four of the six algorithms). In other words, RS is not only a typical case of hyperarticulated speech, but it can also help learning (despite its lack of variability).

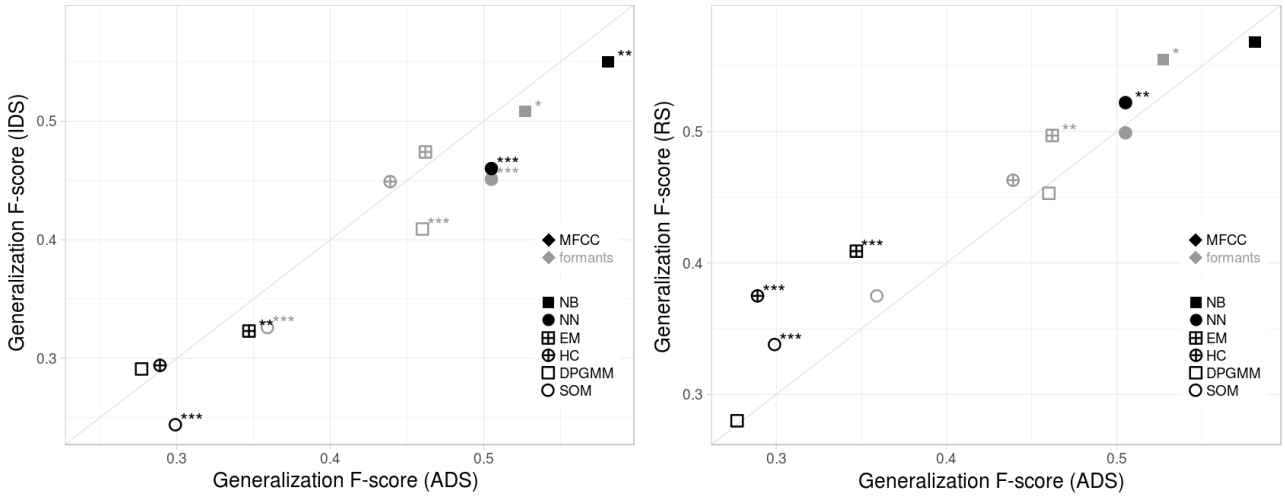
General discussion

Compared to ADS, IDS has been claimed to be simultaneously hyperarticulated (the target categories are farther apart from one another, *e.g.* Kuhl et al., 1997; D. Burnham et al., 2002) and more variable (the tokens of a single category are more distinct from one another, *e.g.* Kirchhoff & Schimmel, 2005; McMurray et al., 2013; Cristia & Seidl, 2014). These two properties, in turn, have been claimed to help phonetic learning for the following reasons: Hyperarticulation makes the categories more separable, hence more easily learnable (Kuhl et al., 1997). Variability helps to build more robust categories, presumably by providing more extreme examples making the categories more distinguishable (Eaves et al., 2016) and enabling to generalize better to novel speakers (Kuhl, 2000). Putting these two properties together would, therefore, attribute to IDS an overall facilitatory effect for robust phonetic category learning. In this paper, we set out testing each of these premises separately, and then exploring their overall predicted effect on category robustness, which we operationalized through the ability of machine learning algorithms to generalize to a novel ADS speaker. We compared the learning performance obtained using ADS and IDS data, with that obtained with RS, a register displaying an increased hyperarticulation, similar to IDS, and a lower variability, such as ADS. By using a register with these characteristics, we attempted to better untangle the effects of hyperarticulation and variability on phonetic learning.

In Experiment 1, we first found in our dataset modest evidence of hyperarticulation in IDS compared to ADS (as measured with a between-category distance), with this effect reaching significance only for the formant representation, but not for MFCCs. Using the same metric, we found a stronger effect for RS which was significantly hyperarticulated when compared to standard ADS, in both formant and MFCC representations. Second, we observed, as expected, that IDS is more variable than ADS, which is itself more variable than RS (all contrasts significant for both representations). Third, when we measured separability, we saw that the two opposite effects of hyperarticulation and higher variability counteracted each other, resulting in a null effect, with IDS not being more separable than ADS.

This is consistent with previous findings on the same dataset, but using different metrics (Martin et al., 2015; Miyazawa et al., 2017; Guevara-Rukoz et al., 2018). Unsurprisingly, we found that separability was strongest for RS, which is both hyperarticulated and less variable. As the hyperarticulation phenomenon does not affect all vowel categories equally (Cristia & Seidl, 2014), a larger effect on the corner vowels might suggest a learning mechanism similar to the one proposed by Adriaans and Swingley (2017), by which the IDS hypearticulated tokens support infants' categorical learning. Employing the formant feature set, we have indeed noticed significantly larger distances between the

[4] The standard deviation across the 100 runs of the three stochastic models was very low (< 0.005 in all cases). An analysis of the results taking into account all 100 runs and other detailed results are presented in the Supplementary Materials, Section S3.



(a) ADS vs. IDS comparison (b) ADS vs. RS comparison

Figure 5. Generalization to novel ADS speakers for the classification of the five Japanese short vowels by six learning algorithms, in either ADS, IDS or RS. Comparisons between ADS and IDS (a) and between ADS and RS (b) respectively, are illustrated when formants or MFCC features were used. The scores represent average F-scores cross-validated on a held out test set of one novel speaker. Displayed next to each point are the p-values of uncorrected paired *t*-tests (*: $p < .05$, **: $p < .01$, ***: $p < .001$). The grey line represents the equal-performance line.

point vowels. It remains to be seen whether the marginal advantage observed for these vowel classes over all the classes is exploited by infants and whether it has an impact on the whole learning process. Further experimental studies would be required to test such hypotheses.

In Experiment 2, we directly tested the claim that inter-speaker variability during learning can be beneficial for category robustness. Specifically, we manipulated the number of speakers in the training set, reasoning that, all other things equal, more speakers during training should yield more speaker-robust categories. This was done by means of six machine learning algorithms covering a large range of possible theories of category learning (from supervised to unsupervised, with Gaussian categories or not). Robustness was measured by generalization to a novel speaker. We found that although some supervised or partially unsupervised learning algorithm benefit from increased speaker variability, fully unsupervised algorithms can be impaired by such variability. Taking into account the fact that, at least initially, the infant's learning algorithms may be unsupervised, it should therefore not be expected that variability is systematically beneficial.

Our results show similar trends to those found in the literature with regard to inter-speaker variability. Even if the differences were not significant for all models, the supervised and partially unsupervised approaches showed a more robust generalization to a novel speaker when trained on data coming from multiple speakers, mirroring the findings of adult (Lively et al., 1993) and infant experimental studies (e.g. Rost & McMurray, 2009; Houston, 2000). Since all the speech materials used in this study represent recordings of mothers interacting with their infant, the results in support of multi-speaker training are also consistent with those obtained for younger infants in Houston and Jusczyk (2000). These studies, including ours, stand in contrast to the conclusions of Kuhl (1979), that low-variability training is sufficient for robust generalization. However, the age of the infants considered in those studies differed, with Kuhl (1979) testing 6-months olds, while the age range in the rest of the studies varied between 7.5 and 15 months. Since different outcomes were obtained with younger infants, it would be appropriate to extend our analyses also to speech addressed to younger infants. Regarding the type of variability, among the experimental studies investigating learning in infants, only Rost and McMurray (2010) has compared both types of variation sources considered here. Similar findings were reported, with intra-speaker variability not helping generalization and inter-speaker variability giving a better generalization (but see Trainor & Desjardins, 2002 for a study showing that intra-speaker variation of another acoustic parameter, pitch range, may help vowel acquisition).

In Experiment 3, we turned to measuring category robustness in different conditions of intra-speaker variability. Our results show that despite being more variable than the other two registers, IDS yields consistently worse, not better, generalization than ADS. This means that the type of within-speaker variability exhibited in the IDS register does not represent a good preparation for the between-speaker variability

exhibited in the generalization tests. However, and somewhat surprisingly, we found that RS *can* be a good preparation for ADS categories despite being less variable. This could point to a possible useful role of book reading in language learning (actually documented in vocabulary development, see Dickinson et al., 2018).

The analysis of our generalization results showed important effects of the model (both supervision type and inductive bias), of the feature types employed, as well as of their interactions (supervision–feature and supervision–feature–inductive bias). This has implications for future computational modelling studies comparing inter-register performance, and also for a better understanding of the outcomes of previous works. Future modelling experiments may take into account the observations made here, such as the importance of input features. Although register differences were not affected by feature type, the latter did interact significantly with both supervision type and inductive bias. Regarding the previous literature, for instance McMurray et al. (2013) and de Boer and Kuhl (2003) used the same type of representations (formants) and obtained very different conclusions – an ADS gain in the former (supervised learning, speech addressed to 9-13-month-olds) and an IDS advantage in the latter (partially unsupervised, speech addressed to 2-5-month-olds). Our results are more in line with the former, although the interaction goes in the direction of the latter (except that, in our case, the partially unsupervised algorithms showed, on average, no difference between ADS and IDS). They are consistent also with other studies employing supervised approaches (Kirchhoff & Schimmel, 2005), indicating and ADS advantage for generalization. Lastly, our results do not reflect those of Eaves et al. (2016), since our DPGMM model employing formant features returned a better performance in ADS than in IDS. However, one must note the contrasting goals of the two studies ("teaching" *vs.* learning) and the subsequent, dissimilar, evaluations (more on this later).

Although we did not find an overall learning advantage for IDS as opposed to ADS, our experiments revealed intriguing patterns when considering the effect of supervision. Supervised models gave overall better results than partially unsupervised models, but this effect was larger in ADS than in IDS. The smaller supervision advantage in IDS might indicate some kind of cognitive advantage – while less could be learned overall, more could be learned without access to the labels. This hypothesis, however, is not supported by the differences between partially unsupervised and unsupervised models. Here again partially unsupervised models fare generally better than unsupervised models, but the difference is stronger in IDS than ADS. This finding may suggest that IDS might be detrimental when trying to infer categories (including their number) directly from the speech signal. Therefore, the final decision on whether IDS is really worse than ADS (unsupervised case) or just as good (partially unsupervised case) may hinge on the availability of other linguistic levels which could provide additional information (Feldman et al., 2009), including the number of phonetic categories of the language (Fourtassi et al., 2014).

In brief, while some of the claims regarding the facilitatory effects of hyperarticulation and variability hold for certain combination of algorithms, register and input representation, the particular mixture of acoustic properties present in IDS addressed to 18-24-month-olds does not, generally, result in a net facilitatory effect as regards phonetic category learning (for a similar account for consonants, see Ludusan, Jorschick, & Mazuka, 2019). If anything, IDS tends to have a small *detrimental* effect across most algorithms. This seems to contradict the idea that the primary function of IDS is to boost language learnability. More analyses are needed to confirm these results on speech addressed to younger infants.

However, our findings do not contradict some of the evidence that hyperarticulation helps language learnability, based on positive correlations between vowel space measures and later language outcome Liu et al. (2003); Hartman et al. (2017); Kalashnikova and Burnham (2018). Indeed, RS, a register exhibiting a high degree of hyperarticulation, gives the best separability and generalizability in our experiments. We found, though, that variability can counteract the beneficiary effect of hyperarticulation. Thus, it would be important that future studies measure both hyperarticulation and variability, in order to be able to disentangle their effects on language outcomes in infants.

In addition, the conclusion that IDS does not help learning might be moderated by the following four considerations.

First, our study is limited by the characteristics of the corpus that we used. Japanese is only one of the many languages in which an IDS register has been documented, and it could be that the acoustic characteristics of IDS and their impact on learnability is language dependent. For instance, English has more vowels than Japanese and some of them display hypoarticulation instead of hyperaticulation (McMurray et al., 2013; Cristia & Seidl, 2014). Although this particular phenomenon would seem unlikely to boost learnability for English IDS, the point remains that the present study should be extended to more languages. Another property of our dataset is that it contains IDS addressed to infants between 18 and 24 months of age, who already have knowledge about the phonetic categories of their native language. While some IDS properties seem to undergo age-related changes (*e.g.* pitch: Stern, Spieker, Barnett, & MacKain, 1983; Kitamura & Burnham, 2003), evidence exists suggesting that vowel pronunciation by caregivers is not modulated by the age of the infant. Longitudinal studies overlapping with the age range of the infants addressed in our study have shown that neither Mandarin (Liu et al., 2009) nor American mothers (E. Burnham et al., 2015) modify the size of their vowel space with the age of the addressee. These results are consis-

tent with the analyses carried out in Experiment 1, showing no age effect on our measures of hyperarticulation, variability and separability. However, it might be that the infants' age could have an effect on the variability and on the degree of category separability in IDS, with speech addressed to younger infants exhibiting different characteristics. Further experimental work is needed to establish such effects. Moreover, although no change in the vowel space of speech addressed to infants was observed, other IDS features that might not be entirely independent from vowel hyperarticulation might be adjusted as infants grow older. These could have influences on the acoustic realization of IDS and the usefulness of these properties for early language acquisition. Therefore, extending our study to new languages and different age groups, especially younger infants, would only help to better establish the generalization of our present findings. In order for this to be done, however, comparable high-quality audio and carefully annotated speech corpora as the RIKEN corpus should be created in other languages and for younger age groups.

Second, IDS affects the whole hierarchy of linguistic structures. Even admitting that IDS has a null or detrimental effect on phonetic category learning, this register could have a positive effect at some other levels (lexical, prosodic, syntactic, semantic), resulting in an overall positive effect on language learnability. To have a fuller assessment of the learnability impact of IDS, it is therefore important to extend the present work to the entire language learning problem. For this, though, computational algorithms able to learn these higher levels of linguistic structures from raw speech should be developed and tested (see Ludusan, Mazuka, Bernard, Cristia, & Dupoux, 2017; Bernard et al., 2020, for some preliminary results on the lexical level and Ludusan, Cristia, Martin, Mazuka, & Dupoux, 2016 for the prosodic level).

Third, Eaves et al. (2016) found that, under certain circumstances, a high variability training set *can* improve unsupervised learning algorithms. This indicates that the detrimental effect of variability that we found in Experiment 3 for unsupervised learning is not a mathematical necessity. Note, though, that Eaves et al. (2016) constructed this high variability training set using strong informational coupling between teacher and learner: in this setting, the teacher monitors the effect of input stimuli on the learner's performance, and adjusts the stimuli accordingly. Even though we found that parental IDS stimuli do not help generic learning algorithms, it could be that each infant has a slightly different learning algorithm – with different weighting of the input dimensions, learning speed, random seed (for stochastic algorithms), etc., for which their parent would provide uniquely tuned IDS stimuli. In other words, parent A could output a specific IDS uniquely tuned for infant A but not for infant B. More research is needed to study this hypothesis, including computational models that estimate the amount and nature of monitoring feedback needed to yield an optimal "teaching" regime and checking this against real data. Incidentally, there is some evidence that parents do modulate their IDS characteristics as a function of the child's linguistic maturity (Newport, Gleitman, & Gleitman, 1977), or based on their feedback (Smith & Trainor, 2008; Lam & Kitamura, 2012) or on their speech perception and processing abilities (Kalashnikova, Goswami, & Burnham, 2018), but that the correlations with measures of child language are not very strong, suggesting some limits on parent's abilities to use this fine-grained monitoring feedback (Newport et al., 1977).

Fourth, even an overall detrimental effect on learnability would not be in contradiction with the fact that infants do pay more attention to IDS than ADS (Cooper & Aslin, 1994; Fernald, 1985; Werker, Pegg, & McLeod, 1994), and that language learning is predicted by the amount of IDS in the environment (Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010; Weisleder & Fernald, 2013). Indeed, IDS has emotional and social qualities (Trainor, Austin, & Desjardins, 2000) which may facilitate learning through increased attention and social motivation (Thiessen, Hill, & Saffran, 2005; Singh, Morgan, & Best, 2002), over and beyond information content and learnability considerations. To take this into account, computational models would have to be equipped with attentional or social filters, instead of assuming that they give equal weight to all input stimuli. This also raises the intriguing possible existence of optimal child-friendly registers combining the learnability benefits of RS and the emotional/attentional benefits of IDS.

To conclude, our study illustrates the general point made in Dupoux (2018) about the importance of computational models run on realistic data, instead of idealized or model-reconstructed data, for shedding light onto unresolved questions concerning infant language development. Such computational studies are a useful complement to experimental studies as they can ascertain the functional role of laboratory measured variables or mechanisms from a learnability point of view. *Vice versa*, such models can also suggest new experiments. For instance, we found that speaker variability can impair certain learning algorithms (supervised algorithms) but help other ones (unsupervised algorithms). This makes the prediction that as infants develop and become more able to exploit top-down information, speaker variability should have a progressively facilitatory effect. Another prediction is that the read speech register should be much more potent than informal registers like IDS or ADS to trigger phonetic learning in infants. All these predictions can, then, be investigated by means of experimental or observational infant studies.

References

Adriaans, F., & Swingley, D. (2012). Distributional learning of vowel categories is supported by prosody in infant-directed

speech. In *Proceedings of the 34th Annual Meeting of the Cognitive Science Society* (p. 72-77).

Adriaans, F., & Swingley, D. (2017). Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *The Journal of the Acoustical Society of America*, *141*(5), 3070–3078. doi: 10.1121/1.4982246

Amano, S., Nakatani, T., & Kondo, T. (2006). Fundamental frequency of infants' and parents' utterances in longitudinal recordings. *The Journal of the Acoustical Society of America*, *119*(3), 1636–1647. doi: 10.1121/1.2161443

Andruski, J. E., Kuhl, P., & Hayashi, A. (1999). The acoustics of vowels in Japanese women's speech to infants and adults. In *Proceedings of the 14th International Congress on Phonetic Sciences* (Vol. 3, pp. 2177–2179).

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *CATESOL Journal*, *30*(1), 177–194.

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, *36*(4), 847–862. doi: 10.1016/j.infbeh.2013.09.001

Bernard, M., Thiolliere, R., Saksida, A., Loukatou, G. R., Larsen, E., Johnson, M., ... Cristia, A. (2020). WordSeg: Standardizing unsupervised word form segmentation from text. *Behavior Research Rethods*, *52*(1), 264–278. doi: 10.3758/s13428-019-01223-3

Boersma, P. (2002). Praat, a system for doing phonetics by computer. *Glot International*, *5*, 341-345.

Burnham, D., Kitamura, C., & Vollmer-Conna, U. (2002). What's new, pussycat? On talking to babies and animals. *Science*, *296*(5572), 1435–1435. doi: 10.1126/science.1069587

Burnham, E., Wieland, E. A., Kondaurova, M. V., McAuley, J. D., Bergeson, T. R., & Dilley, L. C. (2015). Phonetic modification of vowel space in storybook speech to infants up to 2 years of age. *Journal of Speech, Language, and Hearing Research*, *58*(2), 241–253. doi: 10.1044/2015_JSLHR-S-13-0205

Chen, H., Leung, C.-C., Xie, L., Ma, B., & Li, H. (2015). Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *Proceedings of the 16th Annual Conference of the International Speech Communication Association* (p. 3189-3193).

Coen, M. H. (2006). Self-supervised acquisition of vowels in american english. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 2* (pp. 1451–1456).

Cooper, R. P., & Aslin, R. N. (1994). Developmental differences in infant attention to the spectral properties of infant-directed speech. *Child Development*, *65*(6), 1663–1677. doi: 10.1111/j.1467-8624.1994.tb00841.x

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*(4), 913–934. doi: 10.1017/S0305000912000669

de Boer, B., & Kuhl, P. (2003). Investigating the role of infant-directed speech with a computer model. *Acoustics Research Letters Online*, *4*(4), 129–134. doi: 10.1121/1.1613311

Dickinson, D. K., Collins, M. F., Nesbitt, K., Toub, T. S., Hassinger-Das, B., Hadley, E. B., ... Golinkoff, R. M. (2018). Effects of teacher-delivered book reading and play on vocabulary learning and self-regulation among low-income preschool children. *Journal of Cognition and Development*, 1–29. doi: 10.1080/15248372.2018.1483373

Dodane, C., & Al-Tamimi, J. (2007). An acoustic comparison of vowel systems in adult-directed-speech and child-directed speech: Evidence from french, english & japanese. In *Proceedings of the 16th International Congress of Phonetics Sciences* (pp. 6–10).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59. doi: 10.1016/j.cognition.2017.11.008

Eaves, B. S., Feldman, N. H., Griffiths, T. L., & Shafto, P. (2016). Infant-directed speech is consistent with teaching. *Psychological Review*, *123*(6), 758. doi: 10.1037/rev0000031

Englund, K., & Behne, D. (2006). Changes in infant directed speech in the first six months. *Infant and Child Development*, *15*(2), 139–160. doi: 10.1002/icd.445

Feldman, N., Griffiths, T., & Morgan, J. (2009). Learning phonetic categories by learning a lexicon. In *Proceedings of the 31st Annual Meeting of the Cognitive Science Society* (pp. 2208–2213).

Feldman, N., Griffiths, T. L., Goldwater, S., & Morgan, J. L. (2013). A role for the developing lexicon in phonetic category acquisition. *Psychological Review*, *120*(4), 751–778. doi: 10.1037/a0034245

Feldman, N., Myers, E., White, K., Griffiths, T., & Morgan, J. (2011). Learners use word-level statistics in phonetic category acquisition. In *Proceedings of the 35th Boston University Conference on Language Development* (pp. 197–209).

Ferguson, C. (1964). Baby talk in six languages. *American Anthropologist*, *66*(6), 103–114.

Ferguson, C. (1978). Talking to children: a search for universals. In J. H. Greenberg (Ed.), *Universals of human language* (pp. 203–224). Stanford University Press.

Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant Behavior and Development*, *8*(2), 181–195. doi: 10.1016/S0163-6383(85)80005-9

Fernald, A., & Morikawa, H. (1993). Common themes and cultural variations in Japanese and American mothers' speech to infants. *Child Development*, *64*(3), 637–656. doi: 10.1111/j.1467-8624.1993.tb02933.x

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language*, *16*(3), 477–501. doi: 10.1017/S0305000900010679

Fourtassi, A., Schatz, T., Varadarajan, B., & Dupoux, E. (2014). Exploring the relative role of bottom-up and top-down information in phoneme learning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 1–6). doi: 10.3115/v1/P14-2001

Gauthier, B., Shi, R., & Xu, Y. (2007). Simulating the acquisition of lexical tones from continuous dynamic input. *The Journal of the Acoustical Society of America*, *121*(5), EL190–EL195. doi: 10.1121/1.2716160

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development*, *14*(4), 487–513. doi: 10.1016/S0885-2014(99)00016-7

Guevara-Rukoz, A., Cristia, A., Ludusan, B., Thiollière, R., Martin,

A., Mazuka, R., & Dupoux, E. (2018). Are words easier to learn from infant- than adult-directed speech? A quantitative corpus-based investigation. *Cognitive Science*, *42*(5), 1586–1617. doi: 10.1111/cogs.12616

Hartman, K. M., Ratner, N. B., & Newman, R. S. (2017). Infant-directed speech (IDS) vowel clarity and child language outcomes. *Journal of Child Language*, *44*(5), 1140. doi: 10.1017/S0305000916000520

Houston, D. M. (2000). *The role of talker variability in infant word representations* (Unpublished doctoral dissertation). The Johns Hopkins University.

Houston, D. M., & Jusczyk, P. W. (2000). The role of talker-specific information in word segmentation by infants. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(5), 1570. doi: 10.1037//0096-1523.26.5.1570

Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology*, *61*(4), 343–365. doi: 10.1016/j.cogpsych.2010.08.002

Igarashi, Y., Nishikawa, K., Tanaka, K., & Mazuka, R. (2013). Phonological theory informs the analysis of intonational exaggeration in japanese infant-directed speech. *The Journal of the Acoustical Society of America*, *134*(2), 1283–1294. doi: 10.1121/1.4812755

IPA Chart. (2015). http://www.internationalphoneticassociation.org/content/ipa-chart. (Available under a Creative Commons Attribution-Sharealike 3.0 Unported License. Copyright© 2015 International Phonetic Association)

Jansen, A., & Niyogi, P. (2007). Semi-supervised learning of speech sounds. In *Proceedings of the 8th Annual Conference of the International Speech Communication Association* (pp. 86–89).

Kalashnikova, M., & Burnham, D. (2018). Infant-directed speech from seven to nineteen months has similar acoustic properties but different functions. *Journal of child language*, *45*(5), 1035–1053. doi: 10.1017/S0305000917000629

Kalashnikova, M., Goswami, U., & Burnham, D. (2018). Mothers speak differently to infants at-risk for dyslexia. *Developmental Science*, *21*(1), e12487. doi: 10.1111/desc.12487

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238–2246. doi: 10.1121/1.1869172

Kitamura, C., & Burnham, D. (2003). Pitch and communicative intent in mother's speech: Adjustments for age and sex in the first year. *Infancy*, *4*(1), 85–110. doi: 10.1207/S15327078IN0401_5

Kohonen, T. (1988). The 'neural' phonetic typewriter. *Computer*, *21*(3), 11–22. doi: 10.1109/2.28

Kuhl, P. (1979). Speech perception in early infancy: Perceptual constancy for spectrally dissimilar vowel categories. *The Journal of the Acoustical Society of America*, *66*(6), 1668–1679. doi: 10.1121/1.383639

Kuhl, P. (2000). A new view of language acquisition. *Proceedings of the National Academy of Sciences*, *97*(22), 11850–11857. doi: 10.1073/pnas.97.22.11850

Kuhl, P., Andruski, J., Chistovich, I., Chistovich, L., Kozhevnikova, E., Ryskina, V., ... Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science*, *277*(5326), 684–686. doi: 10.1126/science.277.5326.684

Lake, B., Lee, C.-y., Glass, J., & Tenenbaum, J. (2014). One-shot learning of generative speech concepts. In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society* (p. 803-808).

Lam, C., & Kitamura, C. (2012). Mommy, speak clearly: Induced hearing loss shapes vowel hyperarticulation. *Developmental Science*, *15*(2), 212–221. doi: 10.1111/j.1467-7687.2011.01118.x

Liu, H.-M., Kuhl, P. K., & Tsao, F.-M. (2003). An association between mothers' speech clarity and infants' speech discrimination skills. *Developmental science*, *6*(3), F1–F10. doi: 10.1111/1467-7687.00275

Liu, H.-M., Tsao, F.-M., & Kuhl, P. (2009). Age-related changes in acoustic modifications of mandarin maternal speech to preverbal infants and five-year-old children: a longitudinal study. *Journal of Child Language*, *36*(4), 909–922. doi: 10.1017/S030500090800929X

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255. doi: 10.1121/1.408177

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., & Dupoux, E. (2016). Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America*, *140*(2), 1239–1250. doi: 10.1121/1.4960576

Ludusan, B., Jorschick, A., & Mazuka, R. (2019). Nasal consonant discrimination in infant-and adult-directed speech. In *Proceedings of the 20th Annual Conference of the International Speech Communication Association* (pp. 3584–3588). doi: 10.21437/Interspeech.2019-1737

Ludusan, B., Mazuka, R., Bernard, M., Cristia, A., & Dupoux, E. (2017). The role of prosody and speech register in word segmentation: A computational modelling perspective. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 178–183). doi: 10.18653/v1/P17-2028

Martin, A., Igarashi, Y., Jincho, N., & Mazuka, R. (2016). Utterances in infant-directed speech are shorter, not slower. *Cognition*, *156*, 52–59. doi: 10.1016/j.cognition.2016.07.015

Martin, A., Peperkamp, S., & Dupoux, E. (2013). Learning phonemes with a proto-lexicon. *Cognitive Science*, *37*(1), 103–124. doi: 10.1111/j.1551-6709.2012.01267.x

Martin, A., Schatz, T., Versteegh, M., Miyazawa, K., Mazuka, R., Dupoux, E., & Cristia, A. (2015). Mothers speak less clearly to infants than to adults. A comprehensive test of the hyperarticulation hypothesis. *Psychological Science*, *26*(3), 341–347. doi: 10.1177/0956797614562453

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101–B111. doi: 10.1016/S0010-0277(01)00157-3

Mazuka, R., Igarashi, Y., & Nishikawa, K. (2006). Input for learning Japanese: RIKEN Japanese mother-infant conversation corpus (COE Workshop session 2). *IEICE Technical Report*, *106*(165), 11–15.

McMurray, B., Aslin, R. N., & Toscano, J. C. (2009). Statisti-

cal learning of phonetic categories: insights from a computational approach. *Developmental Science*, *12*(3), 369–378. doi: 10.1111/j.1467-7687.2009.00822.x

McMurray, B., Kovack-Lesh, K. A., Goodwin, D., & McEchron, W. (2013). Infant directed speech and the development of speech perception: Enhancing development or an unintended consequence? *Cognition*, *129*(2), 362–378. doi: 10.1016/j.cognition.2013.07.015

Miyazawa, K., Kikuchi, H., & Mazuka, R. (2010). Unsupervised learning of vowels from continuous speech based on self-organized phoneme acquisition model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association* (pp. 2914–2917).

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84–93. doi: 10.1016/j.cognition.2017.05.003

Newport, E., Gleitman, H., & Gleitman, L. (1977). Mother, I'd rather do it myself: Some effects and non-effects of maternal speech style. In C. E. Snow & C. A. Ferguson (Eds.), *Talking to children* (pp. 109–149). Cambridge University Press.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perry, L. K., Samuelson, L. K., Malloy, L. M., & Schiffer, R. N. (2010). Learn locally, think globally: Exemplar variability supports higher-order generalization and word learning. *Psychological Science*, *21*(12), 1894–1902. doi: 10.1177/0956797610389189

Romberg, A. R., & Saffran, J. R. (2010). Statistical learning and language acquisition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *1*(6), 906–914. doi: 0.1002/wcs.78

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, *12*(2), 339–349. doi: 10.1111/j.1467-7687.2008.00786.x

Rost, G. C., & McMurray, B. (2010). Finding the signal by adding noise: The role of noncontrastive phonetic variability in early word learning. *Infancy*, *15*(6), 608–635. doi: 10.1111/j.1532-7078.2010.00033.x

Singh, L., Morgan, J. L., & Best, C. T. (2002). Infants' listening preferences: Baby talk or happy talk? *Infancy*, *3*(3), 365–394. doi: 10.1207/S15327078IN0303_5

Smith, N. A., & Trainor, L. J. (2008). Infant-directed speech is modulated by infant feedback. *Infancy*, *13*(4), 410–420. doi: 10.1080/15250000802188719

Stern, D. N., Spieker, S., Barnett, R., & MacKain, K. (1983). The prosody of maternal speech: Infant age and context related changes. *Journal of Child Language*, *10*(1), 1–15. doi: 10.1017/S0305000900005092

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53–71. doi: 10.1207/s15327078in0701_5

Thiolliere, R., Dunbar, E., Synnaeve, G., Versteegh, M., & Dupoux, E. (2015). A hybrid dynamic time warping-deep neural network architecture for unsupervised acoustic modeling. In *Proceedings of 16th Annual Conference of the International Speech Communication Association* (pp. 3179–3183).

Toscano, J. C., & McMurray, B. (2010). Cue integration with categories: Weighting acoustic cues in speech using unsupervised learning and distributional statistics. *Cognitive Science*, *34*(3), 434–464. doi: 10.1111/j.1551-6709.2009.01077.x

Trainor, L. J., Austin, C. M., & Desjardins, R. N. (2000). Is infant-directed speech prosody a result of the vocal expression of emotion? *Psychological Science*, *11*(3), 188–195. doi: 10.1111/1467-9280.00240

Trainor, L. J., & Desjardins, R. N. (2002). Pitch characteristics of infant-directed speech affect infants' ability to discriminate vowels. *Psychonomic Bulletin & Review*, *9*(2), 335–340. doi: 10.3758/BF03196290

Vallabha, G. K., McClelland, J. L., Pons, F., Werker, J. F., & Amano, S. (2007). Unsupervised learning of vowel categories from infant-directed speech. *Proceedings of the National Academy of Sciences*, *104*(33), 13273–13278. doi: 10.1073/pnas.0705369104

Villa-Vialaneix, N., Mariette, J., Olteanu, M., Rossi, F., Bendhaiba, L., & Boelaert, J. (2018). SOMbrero: SOM bound to realize Euclidean and relational outputs [Computer software manual]. (R package version 1.2-3)

Weisleder, A., & Fernald, A. (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, *24*(11), 2143–2152. doi: 10.1177/0956797613488145

Werker, J. F., Cohen, L. B., Lloyd, V. L., Casasola, M., & Stager, C. L. (1998). Acquisition of word–object associations by 14-month-old infants. *Developmental psychology*, *34*(6), 1289. doi: 10.1037/0012-1649.34.6.1289

Werker, J. F., Pegg, J. E., & McLeod, P. J. (1994). A cross-language investigation of infant preference for infant-directed communication. *Infant Behavior and Development*, *17*(3), 323–333. doi: 10.1016/0163-6383(94)90012-4

Yeung, H. H., & Werker, J. F. (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*, *113*(2), 234–243. doi: 10.1016/j.cognition.2009.08.010